University of Pittsburgh | School of Computing and Information

Dissertation Defense

Doctor of Philosophy in Information Science

“**Geo-distributed Edge and Cloud Resource Management for Low-latency Stream Processing**” by **Jinlai Xu**

Date: November 23rd, 2021
Time: 1:00PM – 3:00PM
Place: https://pitt.co1.qualtrics.com/jfe/form/SV_bJzzFkn8xpotsF0

Committee:

- Dr. Balaji Palanisamy, Associate Professor, School of Computing and Information, University of Pittsburgh
- Dr. David Tipper, Professor, School of Computing and Information, University of Pittsburgh
- Dr. Amy Babay, Assistant Professor, School of Computing and Information, University of Pittsburgh
- Dr. Qingyang Wang, Associate Professor, School of Electrical Engineering and Computer Science, Louisiana State University

Abstract:

The proliferation of Internet-of-Things (IoT) devices is rapidly increasing the demands for efficient processing of low latency stream data generated close to the edge of the network. Edge Computing provides a layer of infrastructure to fill latency gaps between the IoT devices and the back-end cloud computing infrastructure. A large number of IoT applications require continuous processing of data streams in real-time. Examples include virtual reality applications, connected autonomous vehicles and smart city applications. Edge computing-based stream processing techniques that carefully consider the heterogeneity of the computing and network resources available in the infrastructure provide significant benefits in optimizing the throughput and end-to-end latency of the data streams. In an edge computing environment, small-scale micro-datacenters represent ad-hoc collection of computing resources that are geographically distributed. Managing geo-distributed datacenter resources operated by individual service providers raises new challenges in terms of effective global resource sharing and achieving global efficiency in the resource allocation process.

In this dissertation, we propose a distributed stream processing framework that optimizes the performance of stream processing applications through a careful allocation of computing and network resources available at the edge of the network. The proposed approach differentiates itself from the state-of-the-art through its careful consideration of data locality and resource constraints during physical plan generation and operator placement for the stream queries. Additionally, it considers co-flow dependencies that exist between the data streams to optimize the network resource allocation through an application-level rate control mechanism. The proposed framework incorporates resilience through a cost-aware partial active replication strategy that minimizes the recovery cost when applications incur failures. The framework employs a reinforcement learning-based online learning model for dynamically determining the level of parallelism to adapt to changing workload conditions. The second dimension of this dissertation proposes a novel model for allocating computing resources in edge and cloud computing environments. In edge computing environments, it allows service providers to establish resource sharing contracts with infrastructure providers apriori. Based on the established contracts, service providers employ a latency-aware scheduling and resource provisioning algorithm that enables tasks to complete and meet their latency requirements. In geo-distributed cloud environments, it allows cloud service providers to establish resource sharing contracts with individual datacenters apriori for defined time intervals. Based on the established contracts, individual service providers employ a cost and duration-aware job scheduling and provisioning algorithm that enables jobs to complete and meet their response time requirements while achieving global resource allocation efficiency. Based on these mechanisms, we develop a decentralized implementation of the contract-based resource allocation model for geo-distributed resources using Smart Contracts in Ethereum.